Comments on the review of Statistical Inference

Murray Aitkin
University of Melbourne

April 26, 2012

1 Introduction

It is a rare event in the statistical world for a book review to take 15 pages (plus references). The three reviewers are all eminent in the profession, and have tried to be scrupulously careful in their review of a book whose main principle they firmly dispute as non-Bayesian, and which they find not to be useful in their work. The length of the review demonstrates the seriousness with which they regard the book's approach, and the importance to them of denying its relevance to either Bayesian analysis or practical data analysis with complex data.

In responding to the authors' review, I found it helpful to place our different views in the framework of Thomas Kuhn's (1962) discussion of paradigm changes in science.

2 Paradigms

I use the term *paradigms* in the sense of Kuhn: "universally recognized scientific achievements that, for a time, provide model problems and solutions for a community of researchers." (Kuhn 1962 p. viii). My concern is for the relevance of Kuhn's description of the evolution, and revolutionary change, of scientific paradigms to the current state of theories of statistical inference.

In quoting Kuhn at length, I am not endorsing his view of the evolution of paradigms over the views of other philosophers of science; I find his approach and views apposite for the current state of statistics and its theories of inference.

The term paradigm has come to be commonly used in statistics in *the Bayesian paradigm*, and less frequently, in *the frequentist paradigm*. These multiple usages of the term imply, in the Kuhnian sense, that there are separate communities of researchers following these paradigms, since the paradigms are inconsistent in their principles (axioms) for the analysis of data. I will call these separate communities *schools*. This might not matter if these schools attended to different problems in the applications of statistics, but it *does* matter since the applications of the schools overlap very substantially, and the Bayesian and frequentist schools at least claim *universality* in the breadth of their applications.

There are other paradigms (the likelihood paradigm for example), and other approaches to data analysis, notably in survey sampling which I discuss below. I do not discuss the (pure) likelihood paradigm here, as it does not have, or claim to have, the generality of the Bayesian and frequentist schools. There are many practising statisticians who would not regard themselves as members of only one, or any, school. Many statistical scientists are agnostic about theories/paradigms: "I'm not a Bayesian, but I use what works. I'm an empiricist – if MCMC works I use it." (Without a paradigm, how can one tell if anything works?) Statements like these make plain that statistics does not have a *single* paradigm, and the conflict between the major schools is a clear example of what Kuhn calls the *pre-paradigm* state of theories:

> (p. 177) There are schools in the sciences, communities, that is, which approach the same subject from incompatible viewpoints. But they are far rarer there than in other [non-science] fields; they are always in competition, and their competition is usually quickly ended.

3 The Fisherian period

This description fits very well with the pre-Fisherian period (pre-1920), when there was *no* agreed paradigm for statistical analysis. While there were many notable Bayesians (to use the modern term) like Edgeworth, many influential statisticians – Karl Pearson notably – were agnostic in just the sense described above: they used whatever was at hand, (normal) posterior distributions in some analyses, repeated-sampling distributions in others (see John Aldrich's web pages at the University of Southampton Economics Deparment for a stimulating discussion of Pearson and his contemporaries).

Fisher's definition of the likelihood (1912, 1922, 1925) and its theoretical properties for statistical inference was a true revolution. Now inference could be soundly based on a probability model for the data and the likelihood function of the model parameters, without the need for prior distributions on those parameters. The closely related work by Neyman and E.S. Pearson on hypothesis testing had some inconsistencies with Fisher's views, but it further increased the breadth and power of the new paradigm. This paradigm has been immensely successful, and fits exactly Kuhn's description: "a universally recognized scientific achievement that, for a time, provides model problems and solutions for a community of researchers."

4 The Bayesian revival

From about 1990 onwards, Bayesian methods began a major revival. The key to this revival was the development of Markov chain Monte Carlo methods, initially through the Data Augmentation algorithm of Tanner and Wong (1987) which was a Bayesian extension of the immensely important maximum likelihood EM algorithm of Dempster, Laird and Rubin (1977 – this paper had a number

of Bayesian extensions and suggested extensions). This development and its rapidly increasing success illustrates another aspect of Kuhn's argument (pp. 5-6), that the need for a new paradigm develops when the current paradigm seems unable to deal with the increasing complexity of problems, which its own success has brought to the forefront of analysis.

The frequentist difficulties are familiar; for our purposes it is sufficient to point to the major successes of MCMC in complex unbalanced crossed and nested multilevel GLMMS, and the widespread adoption of multiple imputation, with its steady development towards a fully Bayesian analysis with incomplete data.

While maximum likelihood analyses are available for some of these structures, standard errors for the parameter estimates in the ML approach become decreasingly reliable with increasing complexity, and better non-Bayesian precision expressions are almost impossible to obtain.

On the foundational side, the arguments over the reference set for conditional analyses, and those over conditional versus unconditional analyses remain unresolved, and can lead to a wide variety of p-values, as in the ECMO example (A p. 154-60).

5 The state of the Bayesian paradigm

The Bayesian paradigm is now struggling with the frequentist paradigm for model-based supremacy. So far this has left the survey samplers largely unaffected by the dispute, since they generally did not use either Bayesian or model-based frequentist analyses.

The purpose of my book was to strengthen the Bayesian paradigm, and in my view make it a valid and viable *single* paradigm, in two ways: by extending it into the province of the survey samplers, and by plugging a major hole in the current Bayesian paradigm. It is entirely predictable from Kuhn's description of paradigm change, but nevertheless disappointing, that the reviewers overlook or dismiss precisely those contributions which support and strengthen the Bayesian paradigm, while defending, in sometimes distorting and even absurd terms, the difficulties in the current theory as *right*, and even *essential*.

Before dealing with the review, it is necessary for its understanding by readers to explain the nature of the "major hole" in the current paradigm.

6 The Bayes factor

The major objection which frequentists (and others) can raise against the Bayesian paradigm as it is currently implemented is the Bayes factor *anomaly*, to use Kuhn's expression (pp. 5-6):

> Sometimes a normal problem, one that ought to be solved by known rules and procedures, resists the reiterated onslaught of the ablest members of the groups within whose competence it falls.... revealing

> an anomaly that cannot, despite repeated effort, be aligned with professional expectation.

The anomaly arises in the current Bayesian approach to testing point null hypotheses. In the frequentist theory, point null hypothesis testing about a parameter and confidence interval estimation for the parameter are consistent procedures – they are two sides of the same coin. In the current Bayesian theory, testing a point null hypothesis about a parameter through the Bayes factor – a form of generalized likelihood ratio – gives results which may be inconsistent with the credible interval for the parameter: the null value may lie well outside the credible interval, yet the Bayes factor may strongly accept the null value, or find no convincing evidence against it. So Bayes factors and credible intervals are two sides of different coins – their conclusions are incommensurate, as Kass and Raftery (1995, pp. 781-2) point out:

> In frequentist theory, estimation and testing are complementary, but in the Bayesian approach, the problems are completely different.... It may happen that conclusions based on estimation [posterior distribution] seem to contradict those from a Bayes factor. In this case the data seem unlikely under H_0, but if the Bayes factor turns out to be *in favor of* H_0, then the data are *even more unlikely* under H_1 than they would have been under H_0. [authors' emphasis].

How does the anomaly arise? It is a direct consequnce of *integrating the likelihood* $L(\theta)$ *with respect to the prior distribution* $\pi(\theta)$, in models containing an unknown parameter θ for which a prior distribution is available. The use of this approach is justified, in current Bayesian theory, by an appeal to *marginalisation* arguments: the likelihood is regarded as the *conditional probability* of the data given the value of θ, and $\pi(\theta)$ is regarded as the *marginal distribution* of the parameter, so $\bar{L} = \int L(\theta)\pi(\theta)\mathrm{d}\theta$ is regarded as the *unconditional probability* of the data, averaged over θ. This is used to compare models; the ratio of two such integrated likelihoods is called the Bayes factor, and is treated as though it were the ratio of likelihoods from two *completely specified* models.

The particular difficulty which makes this approach anomalous is that it cannot in general use *diffuse* or *non-informative* priors: these have to be *informative* and *parametrized* in terms of some parameter ϕ, which then appears explicitly in the integrated likelihood as $\bar{L}(\phi)$. As the prior becomes more diffuse, the integrated likelihood tends to zero, whatever the data. A consequence of using this approach is that there is much emphasis on *getting the prior right*, which conflicts with the fundamental Bayesian principle that the prior represents one's information about the parameter before the data arrive; it is not something which should be *tuned* to the likelihood to achieve a suitable value of the integral.

This creates a major difficulty for Bayesian statisticians, since they have in principle to decide whether their problem is one of inference about a model parameter through a credible interval, or whether it is a comparison of a null parameter value model with a general model. Having made this decision, the

Bayesian presumably should not look at the other possible analysis – (s)he is not a frequentist.

(A further difficulty, which is not commonly mentioned, is that for regular models with large samples, the $100\alpha\%$ credible interval for the parameter agrees very well with the frequentist $100\alpha\%$ confidence interval, while the Bayes factor conclusion may differ substantially from the frequentist test conclusion.)

How do Bayesians deal with this anomaly? Kuhn describes the framework (p. 37):

> ... [O]ne of the things a scientific community acquires with a paradigm is a criterion for choosing problems that, while the paradigm can be taken for granted, can be assumed to have solutions. To a great extent these are the only problems that the community will admit as scientific or encourage its members to undertake. Other problems, including many that had previously been standard, are rejected as metaphysical, as the concern of another discipline, or sometimes just too problematic to be worth the time.

The Bayesian response is multi-dimensional:

- Adjust the prior: "... to avoid this difficulty, priors on parameters being tested [under the null hypothesis] generally must be proper *and not have too big a spread* ..." (Kass and Raftery p. 782, emphasis added – how big is too big?).
- It is the frequentist theory that is wrong: the p-value from the hypothesis test overstates the strength of evidence against the null model.
- The test of a *precise* null hypothesis is pointless, as we already know the hypothesis is false.
- Problems which are of scientific interest involve well-specified models in which we want to know the parameter ranges, we do not want to compare this model with another model or with a specialised form of this model.
- Model-checking can be done by posterior predictive checks, which compare features of the observed data with those of simulated data sets from the posterior predictive distribution.

These dismissals of hypothesis testing, or model comparison, do not fit easily with the major advantage claimed over the frequentist theory, of being able to compare non-nested models in the same way as nested models, which the frequentist theory cannot.

These responses are reflected in Bayesian textbook treatments of the two-sample and multi-sample normal mean problems, for which the t-test and analysis of variance are the standard tools of the frequentist theory. Almost all these textbooks are silent on the two-sample problem, except for the posterior distribution of the mean difference; there may be references to the overstatement of evidence by the p-value, clearly demonstrated by Berger and Delampady (1987)

and Berger and Mortera (1991) for example. Analysis of variance receives little attention, since this cannot easily be expressed in terms of the posterior distributions of mean differences. There is no Bayesian equivalent of the t-test mentioned.

There is a good reason for this – there were no published Bayesian versions of the t-test until 2005. Now there are four: Aitkin, Boys and Chadwick (2005); Gönen, Johnson, Lu and Westfall (2005); Rouder, Speckman, Sun, Morey, and Iverson (2009); Wetzels, Raaijmakers, Jakab and Wagenmakers (2009).

Some textbooks give a full treatment of the Bayes factor approach, but have to ignore or struggle with its fundamental difficulty. Others dismiss it on the scientific irrelevance grounds mentioned above. Gelman, Carlin, Stern and Rubin (2004) write:

> ... Bayes factors are rarely relevant in our approach to Bayesian statistics... (p. 192)
>
> ... this book has little role for the non-Bayesian concept of hypothesis tests, especially where these relate to point null hypotheses of the form $\theta = \theta_0$ .... most of the difficulties in interpreting hypothesis tests arise from the artificial dichotomy that is required between $\theta = \theta_0$ and $\theta \neq \theta_0$. (p. 250)

7 The review

I will list the reviewers' criticisms and ignored contributions, with relevant comments from Kuhn, and my replies to the reviewers. Quotes from Kuhn are indicated by K, those from my book by A, and others are from the review or are attributed in the reference.

- **The new approach is not useful, and some of the applications given are artificial**

  > Several of the examples in Statistical Inference represent solutions to problems that seem to us to be artificial or conventional tasks with no clear analogy to applied work. (p. 13)
  >
  > For the applied problems that interest us, does the proposed new approach achieve better performances than our existing methods? Our answer, to which we arrive after careful thought, is no. (p. 3)

A key question here is: *what applied problems?* The "simple" (and standard in the frequentist theory) problems of one- and two-group comparisons in A are dismissed as not of interest, and the analysis of a before/after change in attitude (A p.147) is attacked in 1.5 review pages as artificial, the reviewers replacing it by a time series example of their own, which they feel I should have analysed. The data are admittedly artificial, but are relevant to the two correlated group problem I was discussing, not to a time series problem. They were thought

worthy of a Bayes factor analysis by Consonni and La Rocca (2008), which I was comparing with the posterior likelihood analysis. The Bayes factor analysis is not mentioned in the review; but evidently not all Bayesians are of the reviewers' view.

- **The use of improper priors leads to meaningless model comparison procedures**

> When using improper priors lead[s] to meaningless Bayesian procedures for posterior model comparison, we see this as a sign that the Bayesian model will not work for the problem at hand. (p. 2)
>
> ... when one's Bayesian approach leads to a dead end, one must change either one's methodologies or one's beliefs (or both). (p. 2)
>
> ... the other [criticisms of Bayes factors] can be easily rejected on the ground that the posterior distribution of the likelihood is meaningless within a Bayesian perspective. (p. 8)

The word *meaningless* is used four times in the review, referring to either the posterior distribution of the likelihood, or the results from its application. These references give the misleading impression that it is the posterior likelihood comparison of the models which has this meaningless or "dead end" property, when it is actually the current Bayes factor model comparison method that has the meaningless property.

So the authors make the peculiar argument that because their method of choice fails to work for a model with improper priors, while my method works, *I* should *give up the model and the successful method*. A more logical conclusion would be that *they* should *keep the model* but *give up their unsuccessful method*.

- **The use of separate posteriors for the model likelihoods is an error**

> ... the fundamental theoretical argument against using posterior distributions of the likelihoods and of related terms is that the approach leads to parallel and separate simulations from the posteriors under each model. (p.8)
>
> ...the joint posterior on (θ_1, θ_2) [should be proportional to]
>
> $$p_1 m_1(x)\pi_1(\theta_1|x)\pi_2(\theta_2) + p_2 m_2(x)\pi_2(\theta_2|x)\pi_1(\theta_1),$$
>
> [where p_j are the prior model probabilities and m_j are the integrated likelihoods under the proper priors $\pi_j(\theta_j)$.] (p.8)

This curious argument (quoted from Robert and Marin 2008) requires a joint prior distribution *simultaneously on both model parameters*, as in a *mixture* of the two distributions, rather than separate priors for each model's parameter.

Then simulations of parameter values for the likelihoods and likelihood ratio are drawn from their *joint* posterior distribution. While the model comparison or choice problem could be set up in this way, it is quite artificial to do so, since there *is* no mixture, and the two models cannot be simultaneously true. If a further model were to be considered at a later stage, the entire analysis would have to be re-run with a joint prior on all three model's parameters, so that the likelihood ratio distribution *between the first two models* would be altered by the consideration of a third model. This would make model comparisons quite unstable.

If the problem were to decide how many normal components (up to 7) there were in a *real* mixture, as in the galaxy data example discussed below, it would be ludicrous to attempt the analysis with a joint prior on the 77 parameters in all the 7 possible numbers of components, and none of the Bayesian analyses discussed in Aitkin (2011) followed this path.

An accompanying example to this claim (p.9) is of the comparison of a Poisson distribution with a binomial distribution with $n = 5$ trials, given an observed count of $x = 3$. The reviewers show (their Figure 2 p. 9) an "improved discrimination" of the true binomial from their joint prior than from the separate priors. (The Figure caption however describes the true model as negative binomial.)

Since the Poisson is the limiting form of the binomial $b(n, p)$ as $n \to \infty$ and $p \to 0$, it will be no surprise that the binomial is better supported, *if the analyst knows that there are 5 trials.* If not, then n would have to be taken as a second unknown parameter, a much more difficult model comparison problem (Aitkin and Stasinopoulos 1989, see A pp. 25-29).

Does the data analyst know that there are 5 trials? If he does, the answer will be binomial; if he doesn't, why should he compare the Poisson with the binomial with $n = 5$? The example as given has no content.

- **Ignored contributions 1: the galaxy data**

The reviewers make much of the simple one- and two-group examples that they wish to dismiss as irrelevant, while ignoring the complex mixture example. This cannot be because it is not one of "the applied problems that interest us", as Dr Robert is one of the editors of the new Wiley volume on Mixture Applications (Mengersen, Robert and Titterington 2011), in which my 2011 paper appears, which he heard at its presentation in Edinburgh.

This example illustrates both the difficulty of the current Bayesian paradigm, in comparing models with different numbers of normal mixture components, and the simplicity of the comparison through posterior likelihood distributions. The data have been analysed by six different groups of prominent Bayesians, who have produced seven different posterior distributions for the number of normal components in the 82 observations.

These vary widely between remarkably precise (a posterior probability of 0.999 that there are 3 components) and remarably diffuse (ranging fron 3 to 13 components with posterior probability more than 0.01; these posteriors depend

strongly on the parameters in the various prior distributions for the means and variances of the components, because these prior parameters are set – "tuned to the data" to ensure good mixing and convergence of the MCMC procedures used. These settings determine the values of the integrated likelihoods for each model, and therefore the posterior model probabilities.

The posterior likelihood approach by comparison does not depend on prior parameters and can use diffuse priors on the means, variances and proportions in the mixture; it gives zero probability to 1 or 2 components, high probability to 3, low to 4, and probability decreasing rapidly to zero for more than four components.

- **Ignored contributions 2: the recall data**

This example is not mentioned either (Lee 2004, Liu and Aitkin 2008, A pp. 58-63). It has more drastic consequences for the current Bayesian model comparison paradigm, as the author comes to the wrong conclusion about which model is better-supported by the recall data. This is the consequence of the author's trying to be fair to the models: he assigns the same flat prior distribution to the slope and intercept parameters in a set of two-parameter generalised linear one-variable regression models. The problem is that the models with non-linear link and variable transformations have very different likelihood contours from each other and from the linear model, so the integrations over the same flat priors have very different effects on the integrated likelihoods. In particular, the best-fitting (log) model in terms of maximized likelihoods has the tightest likelihood contours, and also has the smallest integrated likelihood, while the hyperbolic model with the highest integrated likelihood has quite diffuse likelihood contours.

The author notes this inconsistency, but gives an extraordinary explanation of it:

> It is also clear, however, that the good fits of the logarithmic and power functions occur at a very narrow set of parameter values, while the hyperbolic function fits reasonably well at a large number of parameterizations. These differences in complexity are even clearer in [his] Fig. 2, which shows contour plots of the same information. [These plots are also given in Liu and Aitkin and in Aitkin (2011).] It can be seen that the hyperbolic function fits across a broader range of parameter values than the exponential function, which in turn has a broader range than the linear, power and logarithmic functions. ... In terms of the balance between goodness-of-fit and complexity, these results indicate that the power and logarithmic models are capable of achieving better fit to the data, but are more complicated than the hyperbolic and exponential functions, because their fit is less robust across parametric variation.

If taken to the extreme, this argument would conclude that a model with a flat likelihood over the whole parameter space would be preferred to a model with

very sharply defined likelihood contours and much higher maximized likelihood, if the prior average of the flat model likelihood over the parameter space was greater than that of the sharply defined model likelihood.

The reviewers say nothing about these examples, but readers of the book will be able to draw their own conclusions on whether "the proposed new approach achieve(s) better performance than our existing methods."

- **Ignored contributions 3: the multinomial/Dirichlet approach to survey sampling**

Chapter 4 of A gives a detailed extension of the multinomial model and Dirichlet prior to stratified and clustered survey designs. This model/prior combination is able to provide a fully Bayesian analysis *without an approximating model for the response distribution.* This is a very important enlargement of the Bayesian paradigm to survey sampling, which has lacked a general theory of analysis since its foundation. The only comment on this work (p.6) is

> This [argument against the use of improper priors] is further illustrated by the use of Haldane's prior in Chapter 4 of *Statistical Inference*, despite it not allowing for empty cells in a contingency table (Jeffreys, 1939)

– a completely unrelated application.

The reviewers presumably do not find survey sampling an applied problem which interests them. It is however the bread, butter and meat of at least one-third of the world's statisticians, especially those in official statistics offices.

- **The inconsistency of the reviewers**

A confusing aspect of the review for any reader is its inconsistency. The reviewers come to overall negative conclusions about the value of the book to them, but these are so hedged about with qualifications that it appears that the reviewers are themselves in some disagreement. I have listed the major criticisms above, and my responses to them, but it is important to list the *approving* comments of the reviewers; these are similar in many respects to the positive reviews by Welsh (in the Australian and New Zealand Journal of Statistics 2011) and Hand (in the International Statistical Review 2011), which readers of this review might also like to read.

> *Statistical Inference* begins with a crisp review of frequentist, likelihood and Bayesian approaches to inference ... (p. 3)
>
> As argued by the author (Chapter 2, page 21), this "small change" in perspective has several appealing features:
>
> - The approach is general and allows to resolve the difficulties with the Bayesian processing of point null hypotheses, being defined solely by the Bayesian model associated with $L(\theta; x)$;

- The approach allows for the use of generic noninformative and improper priors, again by being relative to a single model;
- The approach handles more naturally the "vexed question of model fit", still for the same reason;
- The approach is "simple."(p.4)

Surely this is a strongly positive statement!

> From a frequentist point of view it is of interest to see that the posterior probability of the likelihood ratio being greater than one is approximately a p-value ... in the case of embedded models and under proper priors. This p-value can then be given a finite-sample meaning ... however it seems more interesting from a frequentist perspective than from a Bayesian one. (p. 11)

Contrast this with

> ... the posterior distribution of the likelihood is meaningless within a Bayesian perspective. (p.8)

Continuing with approvals:

> As an evaluation of the ideas found in *Statistical Inference*, the criticisms found in this review are inherently limited. We do not claim here that Aitkin's approach is wrong per se, merely that it does not fit within our inferential methodology, namely Bayesian statistics, despite using Bayesian tools. ... It may thus very well be that the approach of comparing posterior distributions of likelihoods could be useful for some actual applications, and perhaps Aitkin's book will inspire future researchers to demonstrate this. (p.3)
>
> ... once *Statistical Inference* has set the principle of using the posterior distribution of the likelihood ratio (or rather of the divergence difference since this is at least symmetric in both hypotheses), there is a whole range of outputs available including confidence [should be credible] intervals on the difference, for checking whether or not they contain zero. (p.11)
>
> Aitkin makes valuable points – known, but not well-enough known – about the difficulties of Bayes factors, pure likelihood, and other superficially attractive approaches to model comparison. (p. 15)

8 Who owns Bayesian theory?

I conclude with this question, which returns my discussion to Kuhn's book. The reviewers appear to regard themselves as spokespersons for the official Bayesian church, endowed with the power and the authority to determine what is Bayesian and what is not:

> ... we come to Aitkin's book not with a perceived need to rebuild but rather with a view toward strengthening the potential shakiness of the pillars that support our own inferences.(p.2-3)
>
> From our (Bayesian) perspective, this solution [the posterior distribution of the likelihood ratio] (a) is not Bayesian for reasons exposed above, (b) is not parameterization invariant, and (c) relies once again on an arbitrary confidence level.(p.11)

Even the title of the review pushes the message: this is not a Bayesian approach, but a *non-Bayesian* likelihood one.

But Bayesian analysis is not a religious doctrine, it is a *paradigm* – a framework of axioms and procedures which Bayesians have developed, and continue to develop, to serve the purposes of good statistical analysis. Its MCMC development has been remarkably successful.

However, the "axiom" which underlies the integrated likelihood is not serving us well (Bayesians or anyone else), when it integrates the likelihood – the sample evidence for the model – with respect to the prior: the personal calibration of the model parameter in advance of the data.

The "objective" Bayesian school has consistently argued that there should always be available a *reference* analysis with a *neutral* or *noninformative* (relative to the likelihood) prior, to allow us to see *what the data say*, independently of the prior. The integration of the likelihood with respect to the prior takes away that possibility, and replaces it with a nightmare of inconsistent conclusions from varying priors, personal or "data-tuned", as is clear from the galaxy data example. Using the same prior for all models is no solution, as is clear from the recall data example, which raises another spectre – of giving wrong answers in scientific model comparisons problems. Bayesian theory, and Bayesians, will be discredited if this becomes widespread.

I would argue contrary to the reviewers, that the integrated likelihood, as a summary of the evidence for a model, violates another axiom of Bayesian theory, that inference about *any function* of data and model parameters should be through its posterior distribution, not through a plug-in or one-point estimate as in empirical Bayes estimation, which has been consistently criticised by Bayesians for overstatements of precision. The appropriate representation of the information about the model likelihood should be *its* posterior distribution.

Arguments over axioms are part of the development of paradigms, so it is no surprise that the developments I propose have led to major arguments. The development of the Bayesian paradigm is not served, however, by ignoring the realities of its difficulties, the failure of the current solutions to them, and the success of the posterior likelihood approach to them.

9 References

Aitkin, M. (2011) How many components in a finite mixture? In Mengersen, K.L., Robert, C.P. and Titterington, D.M. (eds.) *Mixtures: Estimation*

and Applications. New York: Wiley.

Berger, J.O. and Delampady, M. (1987) Testing precise hypotheses. *Statistical Science* 2, 317-335.

Berger, J. and Mortera, J. (1991). Interpreting the stars in precise hypothesis testing. *International Statistical Review* 59, 337-353.

Consonni, G. and La Rocca, L. (2008) Tests based on intrinsic priors for the equality of two correlated proportions. *Journal of the American Statistical Association* 103, 1260-1269.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39, 1-38.

Fisher, R.A. (1912) On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* 41, 155-160.

Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A* 222, 309-368.

Fisher, R.A. (1925) Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* 22, 700-725.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) *Bayesian Data Analysis*. Boca Raton, CRC Press.

Kass, R.E. and Raftery, A.E. (1995) Bayes factors. *Journal of the American Statistical Association* 90, 773-795.

Kuhn, T.S. (1962) *The Structure of Scientific Revolutions*. University of Chicago Press.

Lee, M.D. (2004) A Bayesian analysis of retention functions. *Journal of Mathematical Psychology* 48, 1-40.

Liu, C.C. and Aitkin, M. (2008) Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology* 52, 362-375.

Mengersen, K.L., Robert, C.P. and Titterington, D.M. (2011) *Mixtures: Estimation and Applications*. New York: Wiley.

Robert, C.P. and Marin, J.-M. (2008) Some difficulties with some posterior probability approximations. *Bayesian Analysis* 3(2), 427-442.

Tanner, M. and Wong, W. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82, 528-550.